



THE NEW SEEKERS

Finding the right thing on the web isn't always easy. Karl Wright explains how to enhance your online investigations and make your site easier for others to spot

By November last year, there were over 100 million unique websites on the internet, and that's not counting other items such as blogs and MySpace pages. That's a lot, and if even just a fraction contain useful information you might actually want, finding those pages among the rest can be frustrating and difficult. Here we show you how to find exactly the results you want, and how you can ensure your own site is easier to find.

WEB DIRECTORIES

When the web was first made accessible to the public in the early 1990s, it was still mainly used by academics to exchange research and talk to their peers. Even at this early stage, however, there was a great deal of uncatalogued information online. If you didn't already know where a particular site or newsgroup was hosted, the chances of finding it were very slim.

In keeping with the enthusiastic 'have-a-go' ethos of the early internet, most of the early solutions to this problem were developed by students working in their own time. Web directories were among the first attempts to

bring order to the internet. These were simple lists of websites, grouped by category. There are considerable disadvantages to this approach. Finding the right information from a list of websites can be a slow, hit-and-miss process. Keeping a web directory up to date enough to be useful is also incredibly time consuming.

One of the best-known web directories of the early internet was Yahoo.com. Originally created by PhD students David Filo and Jerry Yang as a way of keeping track of their own favourite sites, the directory became so huge that the two students had to give up their studies in order to maintain it. Given that both men are now billionaires, they've probably managed to reconcile themselves to those lost doctorates.

SEARCH ENGINE EVOLUTION

At the same time as the early web directories were being created, other groups of students were developing search engines for the web. The very first search engine actually predates the World Wide Web. Called Archie, it was developed by Barbadian computer science student Alan Emtage to search for information

in public FTP archives. Wandex and Allweb, the first search engines designed for the web, appeared in 1993. Unlike their modern counterparts, however, these early search engines looked only at the words in the website's URL. The first search engine to index all the words on a website was Webcrawler, launched in 1994. Within a year, several other search engines had been launched, including the now well-known AltaVista and Lycos. If it works well, a search engine will help you find the information you need more quickly than a web directory.

Although the science of web searching can be tremendously complicated when you examine it in detail, the basics are simple. Behind every search engine is a database of websites. Every site in that database is matched to a number of search criteria, things such as words taken from the page titles and headers, or words that are repeated many times in the page's body text. In order for the search engine to be useful, the database must be updated continually. This is done by programs called bots, or spiders. These programs download



The net closes in The hidden web

Although they are more efficient than web directories, the search engines that most of us use still give us access to only a small portion of the web. In his book *The Invisible Web*, webmaster Chris Sherman estimated that only around 16 per cent of all the information on the web is accessible using search engines such as Google or MSN Search. However, before you get excited by the prospect of hidden information or secret online worlds, it's important to understand what this information is and why it's tricky to find.

Search engines find sites using programs called spiders or bots that 'crawl' from one link to the next. Given that even the most unloved blog or Geocities site is usually linked to by at least one other site, this is a good way of discovering most types of websites. However, if there are no links at all leading to a site, the bots will never discover it.

Similarly, pages that are generated dynamically as a result of a user query to an online database cannot be found using a normal search engine. Most of these hidden pages are part of publicly accessible (if not easily searchable) private and government databases. Much of the information held in these databases may not be that interesting to most people. However, given the amount of information that is available, some of it must be worth reading, particularly if you are an academic, a conspiracy theorist or a journalist.

Luckily, some pages that were previously invisible are no longer so. For instance, search engines can now index PDF documents in the same way they would an HTML page. The invisible portion of the web may be about to shrink even further. Google has filed a patent for a technology that would allow it to search pages that usually require a user to create a login and password before they can be accessed (US Patent application 20060230033). Until now, these pages have defeated search engine bots, preventing search engines from indexing them.

If, however, you are already bursting with curiosity and you can't wait the time it will take Google and its patent, don't despair. The means to search at least some parts of the deep web already exist.

A good place to start is www.completeplanet.com, a deep web search engine created by the IT consultancy Bright Planet, which coined the term 'deep web'. Complete Planet allows you to search on several topics, from gardening to jobs and careers. Its sources are less secret than simply obscure, but it frequently unearths fascinating sources of information. Another good tool is www.turbo10.com, a metasearch site that queries a number of subject specific search engines, all of which have indexed content that normal search engines ignore (often content that isn't even web-based, such as that which resides on peer-to-peer networks).

websites, just as your browser would. Each site is checked against the search engine's index. If it is new, it's added to the index. If it is already in the index, the information associated with the site is refreshed. While this is happening, the bot gets to work on all the links from the downloaded site to other sites. In this way, new sites are discovered and the information about old sites is kept up to date. One of the big differences between search engines is the size of their databases and how efficiently they keep their databases up to date.

When you enter a search string, such as the words 'computer shopper magazine website', your search engine will look in its database for sites containing those words. Enter this search string into Google, and the search engine will find over one million matching websites. If these were presented to you in no particular order, you might never find the real *Computer Shopper* website. To get around this problem, search engines try to assess the relevance of each page they find, and present the most relevant ones to you first. Another big difference between search engines is how they decide which of the pages that match your search string are the most important.

The criteria by which a search engine determines which pages to present first vary.

Some engines look at where, and how often, the words in your search string appear in the document. Increasingly, search engines are following Google's example and assigning each site a 'rank'. This rank is calculated mainly by looking at how many other, unrelated pages link to that page. The more inbound links there are, the more authoritative the search engine assumes the page to be, and the higher the rank it gives it; see the 'Rank and file' box on page 126 for a fuller explanation of Google PageRank. When you search, the engine displays the pages that match your search criteria from highest to lowest.

As we will explain, there are plenty of ways in which you can help search engines to even greater precision, reducing the number of irrelevant entries and increasing the chance that

the exact information you want will be on one of the pages close to the top of your list of search results.

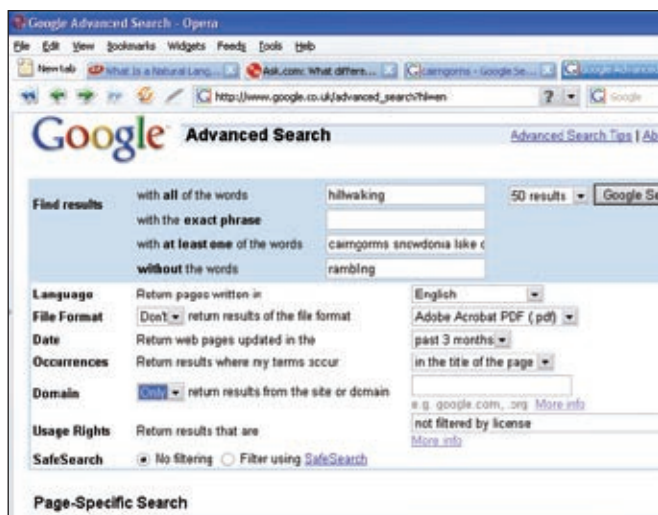
In the past few years, search engines have appeared that don't have their own database of websites matched to keywords, such as Dogpile (www.dogpile.com). Instead, these types of search engines transmit your search request to a number of other search sites, allowing you to use several search engines simultaneously. These are called meta search engines.

SIMPLE SEARCHES

Once you know how a search engine works, you're a step closer to being able to find exactly the results you want in less time. However, before we look at advanced search tips, it's worth going over the basics.



◀ Meta search engines such as Dogpile.com provide you with one interface to search many different search engines. Because you are searching many engines at once, you cannot use site-specific Boolean operators



▲ Most search engines have an advanced search function. Sites often provide a section of tips, too, helping you further refine your search strategy



▲ Google ignores the Boolean operator AND and it does not recognise capitalisation. As you can see from our image search, however, it knows not to split up a good team, searching on all the words you enter even without AND

First, think about how the information you are looking for might be described by someone who knows the topic well. You should try and avoid, or at least qualify, words or search strings that may have double meanings. For instance, if you're looking for the internationally recognised and definitive rules of football, the search string 'football rules' might not be specific enough. You may simply find fan sites written by people who think football is fantastic. However, if you used the string 'FIFA laws of the game' you'll get what you want first time.

If you find that a search string is returning only one or two pages that contain relevant information, don't give up. Browse those pages for alternative words or phrases that might be of more use. Similarly, try using synonyms. So if you can't find what you need searching for "advantages clean fuel" try "benefits clean fuel". Where appropriate, you should try both British and American spellings, though some search engines (notably Google) are smart enough to search both spellings simultaneously.

ADVANCED SEARCH TOOLS

One of the easiest ways to make your search more precise is to use the advanced search tools that are provided by most search engines. These allow you to narrow your search according to several common criteria, such as language, date, format (HTML, PDF and so on) and where in the site the words in your search string should appear (URL, title, body text and so on). Advanced search tools generally allow you to specify whether you want to search for the exact phrase you entered, or all the words but in no particular order, or at least one of the words, and so on. You can usually find a link to a site's advanced search tool somewhere at the top of the page near the search box itself.

SEARCH OPERATORS

The only drawback to using a site's advanced search tool is that it's a bit of a hassle and adds extra time to your search. A far quicker way of doing things is to learn search operators. These are words or signs you can type directly into the search box that narrow and qualify your search in much the same way that a search engine's advanced search tool does.

The search operators that get the most attention are Boolean search operators (named

after the mathematician George Boole). The main Boolean search operators are AND, OR, NOT and sometimes NEAR. Using these operators you can, theoretically at least, define exactly how you want your search terms to be used. Say you simply searched for:

Starsky Hutch

Your search engine would find all pages containing one or both of these words. However, say you searched for:

Starsky AND Hutch

Then your search engine would return only pages that contained both words. If you wanted to find results about someone called Starsky who had nothing to do with the 1970s TV show, you would search using the string:

Starsky NOT Hutch

Search using the NEAR operator, and your search engine will return pages in which your terms stand close to each other in the text.

This may sound great, but a word of caution is necessary. Most search engines don't use

precisely these operators any more. Google, for instance, does not recognise capitalisation or the AND operator. By default, it returns only pages that contain all your search terms. To search for pages that contained individual search terms, as well as pages containing all of them, you would have to search for:

Starsky OR Hutch

What's more, most search engines now have their own set of operators. These often allow you to be very specific in your search criteria, for instance by narrowing your search only to pages that have your search string in their title or the text of their URL.

Don't assume, however, that an operator for one search engine will work in another. For example, if you want to use AltaVista to search for pages with the word Pentium in the title, you would enter the search term:

title:Pentium**

In Google, the same search term would draw a blank. Instead, you should use:

intitle:Pentium

Mind your language Natural language search engines

As an antidote to Boolean search operators, wildcards and other complexities of search, some search engines are pioneering something called natural language search. Instead of cryptic search strings, you type a normal question into the search engine. Yahoo! answers works on this principle, as did the search engine www.askjeeves.com, before it became just www.ask.com.

A Silicon valley company called Powerset has teamed up with Xerox to develop a natural language search application that, the two companies claim, will be as effective as the technology behind Google. Supporters of the natural language search claim that it's easy to use and that it allows people to be precise in what they're searching for in a simplified way. Natural language search could, for example, solve the difficulties caused by

the way in which current search engines ignore common words, such as prepositions. The search strings "pictures of college girls" and "pictures by college girls" should return very different results. As things are, the search engine will ignore the prepositions 'of' and 'by' and, in both cases, provide a list of results sure to make anyone innocently interested in up-and-coming female artists blush.

Detractors point out that most searches at the moment are only a few words long and that you can get precise results without using complex semantics. Anyway, people are used to the way things work. Why would millions of search users want to re-educate themselves in order to use a new product that does the same job as those that they already use and with which they are seemingly quite satisfied?



INTERNET SEARCHING

Our advice in this case is to find a search engine you like, and learn its most important operators. A list of these, or guidance on using them, will be included on the search engine's website. For example, in Google go to Advanced Search then click on the Advanced Search Tips link and scroll down the page.

Before we leave Boolean search, there is one important point to cover. Many search engines such as Google and Yahoo! ignore small, common words and numbers in search strings. So if you search for Jaws 3, the search engine might ignore the number 3. This is where the Boolean operator AND comes in handy. Using it or the relevant equivalent (most commonly, the + sign), you can force the search engine to search on all the words in your search string.

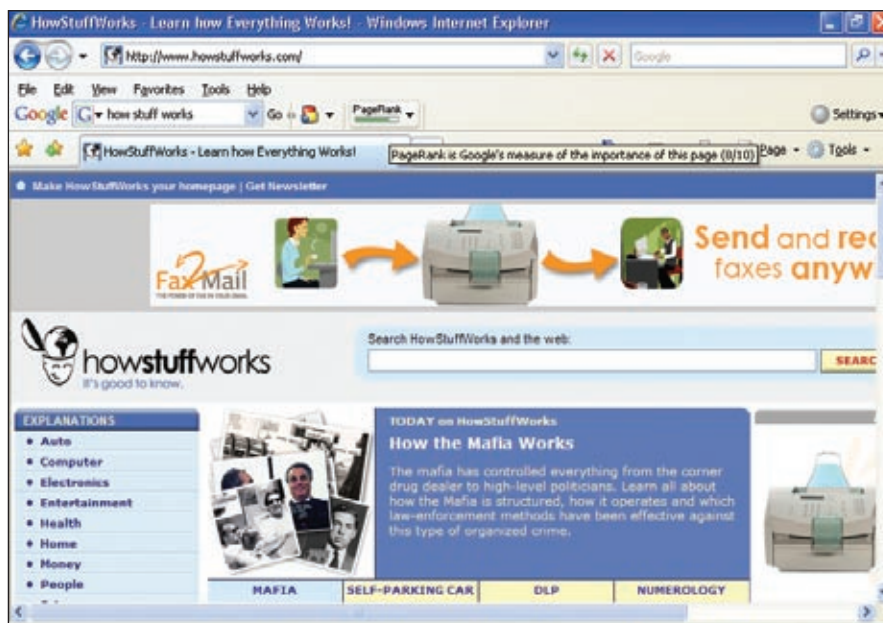
WILDCARDS

Another element of search function that gets a lot of attention is wildcards. Wildcards are symbols that stand in place of unknown characters in your search string. For example, the asterisk is often used in place of an indeterminate number of unknown characters. So if you search for Aber*, your results would include pages for Aberdeen, Aberystwyth and Aberfeldy. If you want to specify exactly the number of unknown characters then you would use question marks, thus Aber???? should return results only on Aberdeen.

However, there's a catch. Wildcard support is by no means universal or consistent. Google uses asterisks, not to indicate a missing portion of a word, but to indicate one or more missing words. Many other engines don't seem to support the wildcards that you might think of as fairly common. Don't worry too much about this. Many search engines are now smart enough to search on the stem of a word, removing the need for wildcards.

SEARCH ENGINE TOOLBARS

It's useful to have search facilities instantly to hand. Nowadays, that's relatively simple as most current browsers have a search box built in. In the Firefox and Opera browsers, this points by default to Google.com. In Internet Explorer 7 the default search engine is, as you'd expect,



▲ Google's toolbar lets you see the PageRank of the page you're on. It also provides instant access to some of the newer features of the Google website, including Gmail, Spellchecker and Google Translation

MSN Search. All the browsers allow you to change the default search provided.

Alternatively, you could use a search toolbar program. Most search engines provide one, prominently offering these free utilities on their search websites. Before search functions were built into most browsers, this was the only way to get an integrated search tool. As well as providing basic search functions, toolbars often block pop-ups, highlight search terms in the webpage's body copy and so on. However, most browsers now perform these functions as well. If you have an up-to-date browser, there aren't many reasons to install a third-party search tool. There is one possible exception. The Google toolbar tells you the page ranking of any given page. This can be useful for anyone with their own website, and we'll come back to this shortly.

WEBSITE OWNERS

Almost everyone who has their own website, unless it's a private site for a club or for family

members, wants more readers. The easiest way to get more readers is to be one of the top results a user sees when they make a relevant search.

Ask anyone whose small business has its own website, and they'll tell you there are dozens of companies that promise they can help you boost your search engine ratings. Because they feel baffled by search engines, and the web in general, many small businesses gladly hand over this task to someone else. With a little careful thought, however, and some patience, you can do the job yourself.

A BLANK PAGE

The first thing to consider is your site's content. Make it interesting and easy to find, and give it a relevant title. This will appear in the title bar of your browser, and is one of the things to which search engines give precedence when deciding how relevant your site is to any given search. Make sure your title is short, snappy and descriptive of your site's most important

Rank and file How Google works

In most respects, Google works like other search engines, using spiders, an index and so on. However, Google also uses parallel computing and its page-ranking system.

By parallel computing, we mean that Google has many index servers, all of which simultaneously search different parts of its index for the words in your search string. This makes it very fast. The PageRank system isn't as straightforward. No-one knows exactly how PageRank works today. But when Google founders Larry Page and Sergey Brin originally published a white paper on what would become Google, they explained PageRank with the following formula:

$$PR(A) = (1-d) + d \left(\frac{PR(T_1)}{C(T_1)} + \text{manually} \dots + \frac{PR(T_n)}{C(T_n)} \right)$$

In this formula, PR(A) is the page rank of page A, which you want to work out. The letter 'd'

is the dampening factor; we'll come to this shortly. T1 stands for the first page that links to page A, Tn for the last page. So PR(T1) is the page rank of page T1. The value C(T1) stands for the number of links from T1 to other pages.

Let's assume T1 is the only page that links to page A, that T1 has a value of one, and that it links to two pages in total. In their white paper, Page and Brin set 'd', the dampening factor, to 0.85, so we will too. Our sum looks like this:

$$PR(A) = (1-0.85) + 0.85 \times (1/2) = 0.575$$

However, if we lower the dampening factor to 0.5, things look completely different:

$$PR(A) = (1-0.5) + 0.5 \times (1/2) = 0.75$$

T1 is 'giving' some of its page rank to page A, without itself losing any of its PageRank value.

The function of the dampening factor is to control how much of one page's value is transferred to another. The lower the dampening factor, the more of the value is transferred.

But this isn't the whole story. There are problems that this system has to overcome – for instance, the circular argument that occurs when two pages link only to each other. The basic principle, however, is hopefully clear. Each link to a page counts as a 'vote' for that page. The more votes a page has, the higher its PageRank and the closer to the top of a list of Google search results it will appear.

However, not all votes are equal. The less discriminate a page is about who it links to, the more outbound links it has, then the less its votes count. This is important to understand when you are trying to get your own site higher up the Google rankings.

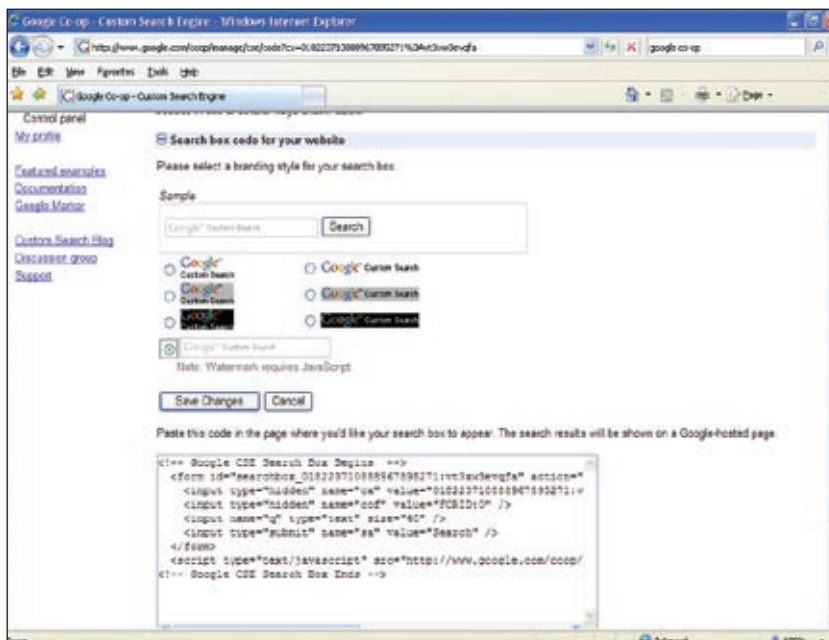


Search party Add a Google search box to your website

To add a search function to your website, you can install a utility on your web server that indexes your webpages and processes any search requests. Microsoft's IIS, the web server included with Windows XP Professional Edition, can do this. If you use a commercial web host, however, you may have to pay to have a search tool added to your site. A good way around this is to use Google's free Custom Search Engine on your own website.

To do this, you must first sign up for an email account with Google. This is called Gmail, also known as GoogleMail in the UK. Once you have Gmail, go to the main Google search page and click on the word 'more', to the right of the search box. Under Explore and innovate, click on Custom Search Engines next to the word heading Co-op. On the next screen click Create Search Engine. Follow the wizard to the end, entering your site's URL when it asks you which sites you want your Custom Search Engine to search.

When you've finished, return to your Google Account homepage and click on My Search Engines. For the search engine you created, click on Control Panel and then, from the list along the top of the page, on Code. Copy and paste the HTML code you find in the next screen into your website's code, and this will add Google's search facility to your website.



▲ As well as adding search functions to your website, Google Custom Search is a nice tool to use in its own right. You could create several Custom Search boxes, each of which searches a number of your favourite websites grouped by category. Paste the code for these into an HTML document that sits on your hard disk, then make that document your browser homepage

content. Put relevant, interesting information on the first page of your site. Never use a blank, or nearly blank, landing page. If the search engine (which looks only at text) can't see anything relevant on your site's homepage, this will probably hurt your page ranking.

You may also have read about metadata, words you can add to the header of your site that a search engine can see but remain invisible to users. Some unscrupulous sites try to misuse this feature to add popular but irrelevant search terms to their site. Don't do this, as you will be blacklisted by search engines. However, putting a relevant meta tag in your site's code can't hurt. Just don't expect it to be a magic fix that will suddenly see your site rocket to the top of the search results.

Adding meta tags to your site is easy. Open the HTML file for your homepage in a text editor such as Notepad and add the meta name tag as follows. Let's say we ran a website about Ford Cortinas and wanted to help other enthusiasts

find our site with a few relevant meta tags. The HTML code would look like this:

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD
HTML 4.01 Transitional//EN"
"http://www.w3.org/TR/1999/REC-
HTML401-19991224/loose.dtd">
<HTML>
<head>
<title> Cortina Cool </title>
<META name="keywords" contents="Ford,
Cortina, Mk I, Mk II, Mk III, Mk IV,
Mk V, 1970s, cars, Taunus, Lotus,
GT, Ford Pinto engine, 4 cylinder,
OHC, Cortina 80">
</head>
```

BUILD IT AND THEY WILL COME

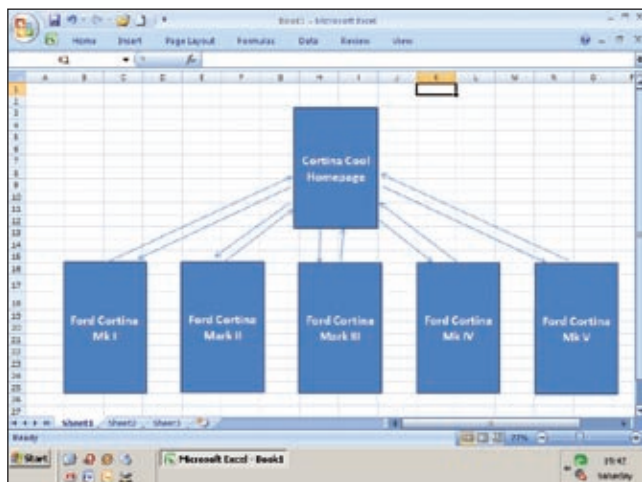
Alongside tailoring your site's content to readers' needs, one of the most obvious (and most often neglected) ways of bringing traffic to your site is to consider the quality of its design and structure.

Unless your audience is particularly arty, it's best to avoid flashy animations and similar features. Make your site easy to use, keep it free of clutter, don't use annoying features such as pop-ups and before adding a design element, ask what benefits it will bring to the reader. It's particularly important that the site is easy to navigate. If your site is easy to use, people are more likely to find the content they want. If they find what they want, they are more likely to come back and more likely to link to you.

The fact that there are more links pointing to your site will help people find you. In addition, in Google each link from another page to you counts as a 'vote' that boosts your PageRank value (see the 'Rank and file' box opposite). Other search engines also now measure a site's inbound links to determine how interesting that site is, so you won't just be boosting your Google rating. However, don't try and cheat the system by creating gratuitous links on other people's sites (for instance, by posting links on blogs or forums). Search engines ignore these links to your site, and may even blacklist you because of them. To see how many pages link to you, type your website's URL into Google preceded by the word link and a colon, for instance:

link:www.computershopper.co.uk.

Finally, remember that Google looks at individual pages, not entire sites; if you run a site on 1990s soft rock groups, don't be surprised if the page for the band Bare Naked Ladies gets the most hits. Links between pages also count towards your page ranking. If your site is quite small, say 20 pages, you can use a sidebar to link every page to every other page. If it's bigger than this, linking all the pages will create confusion. You should make as many relevant internal links as possible; just don't get carried away. **CS**



◀ Carefully plan the structure of your website, flatplanning it on paper before you even start. You want the maximum number of internal links that's practical, but not at the cost of turning your site into an unstructured and confusing mess that's difficult to navigate